THE DESIGN OF CMOS RADIO-FREQUENCY INTEGRATED CIRCUITS

SECOND EDITION

THOMAS H. LEE
Stanford University

CAMBRIDGE
UNIVERSITY PRESS

drain currents, and hence that of the NMOS V_{gs}. Thus, the diode voltage appears across R; the corresponding current is the same in both halves of the mirror and is therefore the bias current of the diode itself. Thus, as in the op-amp version of this circuit, the diode provides its own bias current.

The self-biased circuit of Figure 10.4 is quite versatile. It should be clear that the diode may be replaced by a variety of elements. For example, a diode-connected MOSFET would produce a bias current of V_{gs}/R, or a zener diode (if available) could be used instead. As we'll see in the next section, the self-biased circuit is particularly useful in realizing bandgap voltage references in CMOS technology. Some minor subtleties concerning the operation of the quad of MOSFETs in Figure 10.4 are considered in Problem 12.

10.5 BANDGAP VOLTAGE REFERENCE

Because IC technology directly offers no reference voltages that are inherently constant, the only practical option is to combine two voltages with precisely complementary temperature behavior. Thus, the general recipe for making temperature-independent references is to add a voltage that goes up with temperature to one that goes down with temperature. If the two slopes cancel, the sum will be independent of temperature.

Without question, the most elegant realization of this idea is the bandgap voltage reference. It produces an output voltage that is traceable to fundamental constants and therefore relatively insensitive to variations in process, temperature, and supply.

The first widely used bandgap voltage reference was designed by Bob Widlar in the hugely popular and revolutionary LM309 5-V regulator IC from National Semiconductor. It was the first reference whose initial accuracy was good enough to eliminate the requirement for adjustment by the end user. Thus, only three terminals were needed (allowing use of inexpensive transistor packages), making this part as easy to use as one could hope.

To understand quantitatively how bandgap references work, we need to re-examine the detailed behavior of junction voltage with temperature. Since transistor junctions exhibit more nearly ideal characteristics than ordinary diodes, we will assume bandgap implementations that use transistors. A plot of V_{BE} versus temperature is sketched in Figure 10.5.[6]

Recall that V_{BE} is nearly perfectly CTAT (i.e., it goes down linearly with temperature). Now suppose we add to this CTAT V_{BE} a voltage that is perfectly proportional to absolute temperature (PTAT). If we choose the slope of the PTAT term equal in magnitude to that of the CTAT term, the sum will be independent of temperature (see Figure 10.6). We see that something funny happens above about 600 K, but the fact

---

[6] Again, keep in mind that this plot of V_{BE} is a slight fiction because we have neglected the small curvature caused by the weak temperature dependence of I_0. The correction is second-order, and we will take care of this little detail shortly.

voltage appears
he mirror and is
) version of this

be clear that the
diode-connected
le (if available)
ed circuit is par-
chnology. Some
s in Figure 10.4

FIGURE 10.5. V_{BE} versus temperature.

: inherently con-
ecisely comple-
ng temperature-
ature to one that
. be independent

bandgap voltage
mental constants
ture, and supply.
y Bob Widlar in
National Semi-
good enough to
y three terminals
king this part as

FIGURE 10.6. Illustration of bandgap reference principle.

ed to re-examine
: transistor junc-
, we will assume
is temperature is

ly with tempera-
ctly proportional
AT term equal in
temperature (see
0 K, but the fact

neglected the small
s second-order, and

that the principle fails at temperatures high enough to melt lead is rarely a practical concern.

Note that the addition of a PTAT and CTAT voltage in the proper ratio yields an output equal to the bandgap voltage (extrapolated to 0 K), independent of temperature. Stated another way, if we adjust the PTAT component to make the output voltage equal to V_{G0} at any temperature, then the output voltage will equal V_{G0} at all temperatures — at least in this slightly simplified picture.

At this point, it's natural to consider how one obtains a PTAT voltage, since this whole concept relies on having one around. Let's start with the familiar equation for V_{BE}:

$$V_{BE} = V_T \ln\left(\frac{I_C}{I_S}\right). \tag{5}$$

Using this expression, we can readily compute the *difference* in two V_{BE}s for identical transistors operating at two different values of collector current (or, more generally,
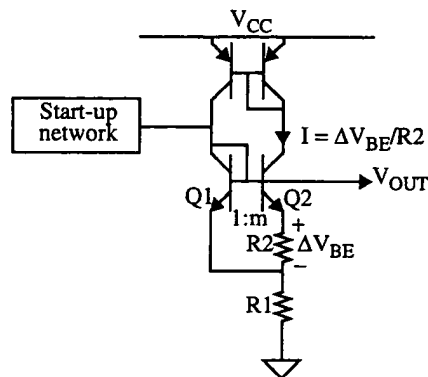
FIGURE 10.7. Classic Brokaw bandgap reference circuit.

for transistors made in the same process, operating at two different values of collector current density):

$$\Delta V_{BE} = V_{BE2} - V_{BE1} = V_T \ln\left(\frac{J_{C2}}{J_{C1}}\right). \tag{6}$$

The misleading I_S term drops out, so we can conclude confidently that ΔV_{BE} truly is PTAT if the collector current densities are in a fixed ratio. Thus, while each V_{BE} is nearly CTAT, the *difference* between two V_{BE}s is perfectly PTAT.

10.5.1 CLASSIC BANDGAP REFERENCE

Now that we've got all the ingredients, all that remains is to sum the CTAT V_{BE} term with the right amount of PTAT ΔV_{BE}. Although one could imagine a number of methods for doing so, the Brokaw cell is a particularly elegant (and accurate) implementation of the bandgap reference. The classic bipolar implementation is shown in Figure 10.7 (again, basic mirrors are shown for simplicity's sake); we'll modify this circuit shortly for implementation in CMOS technology.

As we shall see, the output voltage is the sum of a PTAT voltage and a V_{BE}. Here, Q_1 and Q_2 operate at a fixed current density ratio of m (>1) set, for example, by ratioing the emitter areas. Now, by KVL, the voltage across R_2 is the difference in V_{BE}s of Q_1 and Q_2, and is therefore PTAT and equal to $V_T \ln m$. Assuming that the TC of R_2 is negligibly small, the current passing through it will also be PTAT. Furthermore, the current through R_1 is simply twice that through R_2, since the two collector currents are equal.[7] Therefore, the voltage drop across the entire resistor string is purely PTAT. Finally, the output voltage is just this PTAT voltage plus the V_{BE} of Q_2, as advertised. With proper choice of R_1 and R_2, the output voltage will have zero TC. As a free bonus, a PTAT voltage is available at the emitters of Q_1 and Q_2, providing thermometer outputs.

---

[7] We are neglecting errors due to mismatch, nonzero base currents, and finite Early voltage.

Design Example

To carry out an actual design, we need some characterization data for our process. As a specific example, suppose we go to the lab and find that $V_{BE} = 0.65$ V at 300 K and 100 μA for a transistor of Q_2's size. Furthermore, let $m = 8$. This choice[8] of m sets $\Delta V_{BE} = 53.8$ mV (a number comfortably larger than any offsets that we expect) at 300 K. Since we definitely know the value of V_{BE} at 100 μA, a prudent choice for the collector currents would be this value of 100 μA, and this choice then fixes the value of $R_2 = \Delta V_{BE}/100 \mu$A $= 538$ Ω. Now, since we want the output voltage to be 1.2 V, the drop across R_1 must be $1.2 - V_{BE} - \Delta V_{BE} = 0.496$ V. Finally, noting that the current through R_1 is twice that through R_2, we conclude that we should choose $R_1 = 0.496$ V$/200 \mu$A $= 2.48$ kΩ, completing the design.

You may have noticed that the collector currents in the Brokaw cell are not constant (in fact, they are PTAT if we assume that the resistors have zero TC). To see why this does not invalidate all we've done so far (in fact, it is beneficial), it is now time to take care of a few details – namely, those involving the temperature dependency of I_0.

A quasiempirical expression for I_0 is

$$I_0 = A_E BT^r, \tag{7}$$

where A_E is the emitter area, B is a process-dependent constant, T is the absolute temperature, and r is a process-dependent quantity we'll call the *curvature coefficient*. For the relatively deep, diffused emitters of older bipolar processes, r typically has a value between 2 and 3, while for the shallow, implanted (and very heavily doped[9]) diffusions that are more common in modern CMOS and high-speed bipolar processes, r typically ranges from 4 to 6.

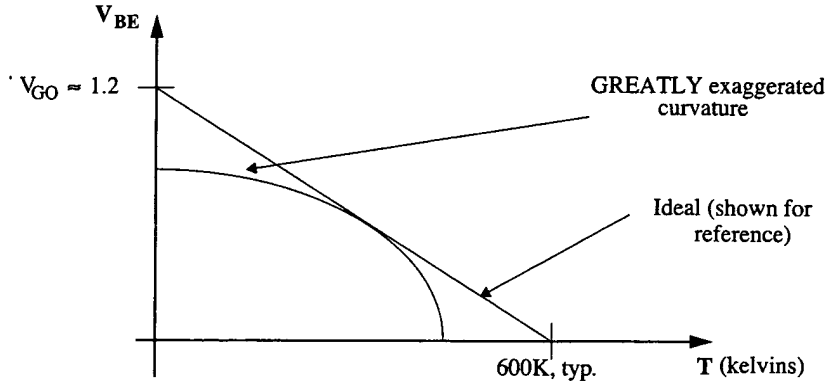With this equation for I_0, we can express V_{BE} as follows:

$$V_{BE} = V_{G0} - V_T \ln\left(\frac{A_E BT^r}{I_C}\right). \tag{8}$$

Plotting in Figure 10.8 as before, we can see why it is reasonable to call the parameter r the curvature coefficient (aside from the euphonious alliteration).

Because the argument of the log is not quite independent of T, the temperature coefficient of V_{BE} is not quite constant, leading to a small departure from CTAT behavior for V_{BE}. Additionally, we've seen at least one implementation of a bandgap reference in which the collector current is also not constant. So, let's compute the actual TC that results if, in addition to the temperature dependence of I_0, we also consider a collector current that varies as the nth power of T:

---

[8] It is important that Q_2 be laid out as eight instances of Q_1 to guarantee that Q_2 behaves as eight parallel devices of Q_1's size. If Q_1 is placed at the center of a common-centroid arrangement, errors due to process variation will be minimized.

[9] Bandgap narrowing and nonlinearity in the heavily doped emitters are probably responsible for the high values of r.

FIGURE 10.8. V_{BE} versus temperature.

$$\frac{dV_{BE}}{dT} = \frac{d}{dT}\left[V_T \ln\left(\frac{CT^n}{A_E BT^r}\right)\right] = \frac{d}{dT}\left[V_T \ln\left(\frac{C}{A_E BT^{r-n}}\right)\right], \tag{9}$$

so that

$$\frac{dV_{BE}}{dT} = \frac{k}{q}\left[\ln\left(\frac{C}{A_E BT^{r-n}}\right) - (r-n)\right], \tag{10}$$

which we may rewrite in a somewhat lower-entropy form, as in Section 10.2:

$$\frac{dV_{BE}}{dT} = -\frac{[V_{G0} - V_{BE} + (r-n)V_T]}{T}. \tag{11}$$

Note that the curvature term disappears if $r = n$, and we're left with the same expression for the temperature coefficient as derived earlier. In the Brokaw cell, $n = 1$, which reduces the effect of yet does not cancel r (remember, r is typically a minimum of 2, and can range up to about 6). Graphically, think of the increasing collector current with temperature as straightening out the V_{BE} curve.

The next question is: How does the curvature term affect the bandgap reference itself? The most expedient answer comes from deriving the condition for net zero TC. Suppose we call GV_T the PTAT component that we add to V_{BE}. Then the TC of the PTAT component may be written as GV_T/T, so the condition for zero TC is

$$\frac{dV_{BE}}{dT} + \frac{GV_T}{T} = 0 \implies G = \frac{[V_{G0} - V_{BE} + (r-n)V_T]}{V_T}, \tag{12}$$

which corresponds to an output voltage of

$$V_{out}\big|_{TC=0} = V_{BE} + GV_T = V_{G0} + (r-n)V_T. \tag{13}$$

This last equation depends on V_T and therefore implies that the output voltage cannot have zero TC at all temperatures; the best we can do is achieve zero TC at one temperature. Furthermore, in order to achieve this zero-TC condition at that one temperature, we need to adjust the output to a voltage *higher* than V_{G0} by an amount equal to $(r-n)V_T$. Fortunately, this correction term is relatively small, typically amounting

Table 10.1. *Output voltage as function of T and r − n*

$r - n$	V_{out} @ $T = -55°C$	V_{out} @ $T = 50°C$	V_{out} @ $T = 150°C$	Maximum error
1	1.226	1.228	1.227	2 mV
2	1.252	1.256	1.253	4 mV
3	1.279	1.284	1.280	5 mV
4	1.305	1.312	1.307	7 mV
5	1.331	1.339	1.333	8 mV

to just tens of millivolts out of a total that is greater than a volt. Hence, the output need only be trimmed to a value a few percent greater than V_{G0} at the temperature where zero TC is desired (generally, the center of the operating temperature range).

At this point, we'd like to quantify the errors that are caused by the curvature. Unfortunately, although the equations we've developed so far are valuable for design, they aren't quite suitable for analysis. To derive one that is, let us choose the factor m so that the output voltage has zero TC at some temperature we'll call T_R (for the reference temperature). Throughout, we'll use the subscript R to denote a variable's value at this reference temperature. With this notational convention, we may express V_{out} as follows:

$$V_{\text{out}}(T) = V_{G0} + \frac{T}{T_R}(r - n)V_{TR} - \frac{T}{T_R}(r - n)V_{TR} \ln\left(\frac{T}{T_R}\right), \tag{14}$$

or, as some prefer,

$$V_{\text{out}}(T) = V_{G0} + \frac{T}{T_R}(r - n)V_{TR}\left[1 - \ln\left(\frac{T}{T_R}\right)\right]. \tag{15}$$

Note that this equation has the right limiting behavior: when $T = T_R$, it yields the output voltage corresponding to the zero-TC condition. Also note that, if we were able to arrange for the collector currents to vary as T^n with $n = r$, then the output voltage would have zero TC at all temperatures if V_{out} were adjusted to a value V_{G0} at any temperature. This last observation is at the core of many efforts to synthesize curvature-corrected bandgap references.

Even without elaborate curvature-correction methods, though, the Brokaw cell (where $n = 1$ in the classic implementation) provides outstanding performance; the curvature inherent in bipolars is simply not all that bad, and the Brokaw cell contributes little error of its own. To illustrate this point, let's compute the actual error one could expect, over a temperature range of −55°C to +150°C, if T_R is chosen as +50°C and if the quantity $r - n$ ranges from 1 to 5; see Table 10.1.

As is evident, the total change in output voltage is less than a percent over the entire temperature range, even with relatively large values of $r - n$. Furthermore, the output is a maximum at the reference temperature, and drops off above and below this temperature in a quasiparabolic manner. As a consequence, setting T_R equal to
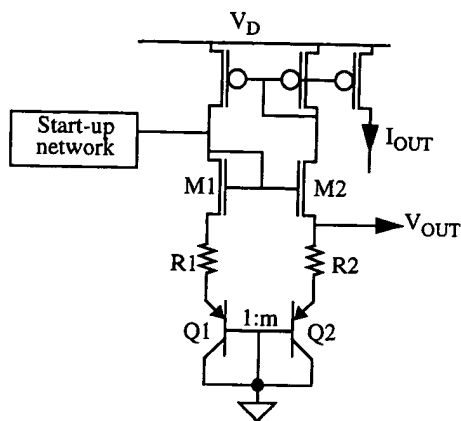
FIGURE 10.9. CMOS bandgap reference.

the center of the desired operating temperature range nearly minimizes the maximum deviation from the value at T_R.

As a final note, the levels of performance in Table 10.1 assume an ideal scenario in which second-order errors (due to device mismatch, nonzero resistor TC, β drift, etc.) are ignored. Actual performance will be somewhat worse in practice owing to the combined effect of these sources. Careful layout of all devices is mandatory in order to minimize errors.

10.5.2 BANDGAP REFERENCES IN CMOS TECHNOLOGY

The classic Brokaw cell uses bipolar transistors in which all device terminals float, so it cannot be implemented directly in this form in CMOS technology. Rearranging to accommodate the restrictions placed on the parasitic substrate p–n–p yields the circuit shown in Figure 10.9.

Transistor Q_2 is designed to have m times the emitter area of Q_1. The quad of CMOS transistors enforces equal emitter currents, so that the *collector* current density ratio is approximately m. Implicit in the last statement is that this circuit has a greater sensitivity to β than the original Brokaw cell. This unfortunate consequence of being forced to use the substrate p–n–p's leads to larger errors than the classic bandgap cell, particularly because β is rarely large enough to be ignored (values of 5–10 are typical). Nevertheless, even a poorly performing bandgap reference is considerably superior to anything that can be built out of pure CMOS components.

Choosing component values for this circuit proceeds in a manner quite similar to that for the classic cell. Begin by specifying a reference temperature T_R at which the TC is to be zero. For illustrative purposes, assume that this temperature is to be 350 K.

The next step is to calculate the target output voltage at this reference temperature. As mentioned previously, the shallow, heavily doped p+ diffusions used to make the emitters lead to relatively large curvature coefficients, with r typically 4 or 5. If no

---

device models a
the quantity r –

Now assume
and that the lar;
is simply

If we assume a

thus completin;
As a final cc
tively constant
offsets the PTA
to the PMOS n
justed through
a bias current i

As a final cc
the relatively ;
errors. Using l
beneficial, but
as switches, al
supposed to n
every alternati
example, then
technique, unt
CMOS proces
plementations

A constant cu
case. Importa
must be held c

A circuit w
a reference re
we've already

[10] V. Ceekala et
Digest of Tec
2001, granted

device models are available, a reasonable starting point is to assume a value of 4 for the quantity $r - n$. Hence, the target output voltage should be

$$V_{\text{out}} = V_{G0} + (r - n)V_T \approx 1.32 \text{ V}. \tag{16}$$

Now assume that we have selected 100 μA for the individual emitter currents and that the larger transistor, Q_2, has a V_{BE} of 0.65 V at this current at T_R. Then R_2 is simply

$$R_2 = \frac{V_{\text{out}} - V_{\text{BE2}}}{I_E} = 6.7 \text{ k}\Omega. \tag{17}$$

If we assume an m equal to 8 then V_{BE1} is about 63 mV larger than V_{BE2}. Hence,

$$R_1 = \frac{V_{\text{out}} - V_{\text{BE1}}}{I_E} = 6.07 \text{ k}\Omega, \tag{18}$$

thus completing the design.

As a final comment on this circuit, one usually finds that the bias current is relatively constant with temperature because resistors typically have a positive TC, which offsets the PTAT tendency of the core design. Thus, currents from a mirror slaved to the PMOS mirror will be roughly constant. The precise TC obtained may be adjusted through a suitable choice of resistor values if the ultimate goal is to generate a bias current rather than a reference voltage.
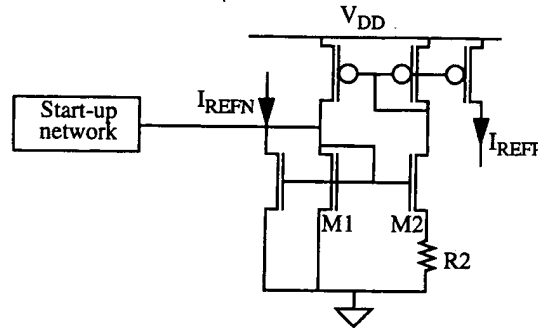
As a final comment on CMOS bandgap circuits in general, it must be noted that the relatively poor matching that CMOS normally exhibits can result in significant errors. Using large devices and operating them at moderately high gate overdrives is beneficial, but it costs power. An alternative technique is to make use of MOSFETs as switches, alternately exchanging the left and right transistors of every pair that is supposed to match.[10] By symmetry, the effect of mismatch should reverse sign at every alternation. If we simply time-average the output with a low-pass RC filter, for example, then the output voltage will be insensitive to offset, to first order. Using this technique, untrimmed 3σ errors of about 1% have been demonstrated in a 0.18-μm CMOS process. Such performance is competitive with many trimmed bipolar implementations. This general technique may be applied wherever symmetry allows it.

10.6 CONSTANT-g_m BIAS

A constant current or constant voltage is often desirable, but this is not always the case. Important examples include situations in which it is the transconductance that must be held constant, such as in the low-noise amplifiers described in Chapter 12.

A circuit whose bias current corresponds to a g_m that is inversely proportional to a reference resistance is a modification of the self-biased CMOS quad of transistors we've already seen; this is shown in Figure 10.10.

---

[10] V. Ceekala et al., "A Method for Reducing the Effects of Random Mismatch in CMOS," *ISSCC Digest of Technical Papers*, February 2002. Also see U.S. Patent #6,535,054, filed 20 December 2001, granted 18 March 2003.

FIGURE 10.10. Basic constant-g_m reference.

To derive an expression for the transconductance of M_1, first use KVL to write

$$V_{gs1} = V_{gs2} + IR_2 \implies V_{od1} = V_{od2} + IR_2, \tag{19}$$

where we have assumed perfect PMOS mirrors and equal threshold voltages for the two long-channel NMOS transistors.

If transistor M_2's width is m times that of M_1, then the two overdrive voltages are related as follows:

$$V_{od2} = \frac{V_{od1}}{\sqrt{m}}. \tag{20}$$

For the long-channel devices we have been considering all along, g_m is simply $2I/V_{od}$. Therefore,

$$V_{od1} = V_{od2} + IR_2 \implies V_{od1} = \frac{V_{od1}}{\sqrt{m}} + IR_2 \implies \frac{2I}{g_{m1}} = \frac{2I}{\sqrt{m}g_{m1}} + IR_2. \tag{21}$$

Solving for the transconductance yields

$$g_{m1} = \frac{2(1 - 1/\sqrt{m})}{R_2}, \tag{22}$$

revealing explicitly that the transconductance is proportional to the reciprocal of reference resistance R_2. If the ratio m is precisely 4, then the proportionality becomes an equality. The currents I_{REFN} and I_{REFP} generated by this cell may be mirrored to slave devices (with or without scaling) to set the transconductance of multiple NMOS devices in other parts of a circuit.

Because this bias generator depends on the quality of the reference, demanding applications may require the use of an external resistance. In noncritical applications, ordinary on-chip resistances (e.g., unsilicided poly) may suffice.

The foregoing development rather optimistically assumes that both NMOS transistors possess the same threshold voltages, despite the unequal source-to-bulk potentials for the devices. In practice, the back-gate bias effect produces unequal threshold shifts and thus introduces an error. To minimize the impact of this back-gate effect, choose relatively high current densities for the devices – within the constraints of maintaining long-channel operation. Doing so assures large overdrives, making the overall
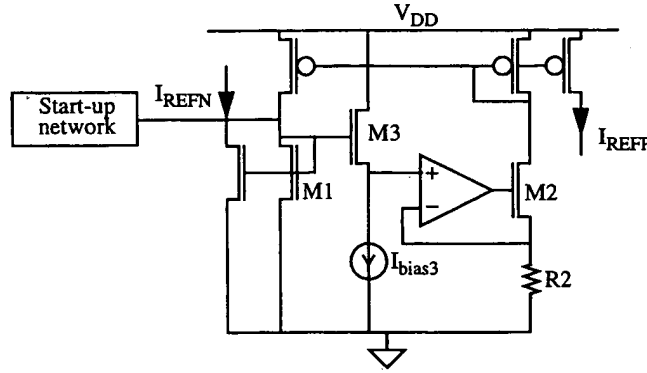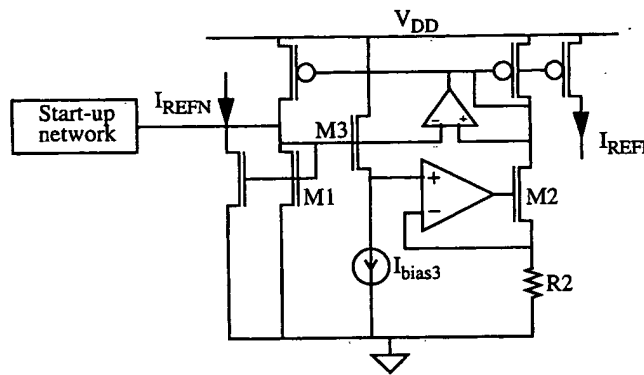
circuit's behav
low IR_2 produ
these strategie:
An alternati
is largely irrel
resistor R_2 equ
M_3. The curre
a level shift ec

$$V_{gs1} -$$

Because of th
this circuit to
IR_2 products.
Mismatche
cuits. Matchi
overdrives. S
close to equal

FIGURE 10.11. Improved constant-g_m reference.

FIGURE 10.12. Minimum-error constant-g_m reference.

circuit's behavior less sensitive to threshold differences. Finally, select a relatively low IR_2 product to minimize the difference in source-to-bulk voltages. Use of both of these strategies together usually permits the attainment of satisfactory performance.

An alternative constant-g_m reference is shown in Figure 10.11. In this circuit, M_2 is largely irrelevant, thanks to the op-amp. As seen in the figure, the voltage across resistor R_2 equals M_1's V_{gs} level-shifted downward by an amount equal to the V_{gs} of M_3. The current I_{bias3} is chosen in conjunction with the dimensions of M_3 to produce a level shift equal to the threshold voltage of M_1. Therefore,

$$V_{gs1} - V_{gs3} \approx V_{gs1} - V_{t1} \approx IR_2 \implies \frac{2I}{g_{m1}} \approx IR_2 \implies g_{m1} \approx \frac{2}{R_2}. \quad (23)$$

Because of the additional degree of freedom represented by I_{bias3}, it is possible for this circuit to exhibit smaller errors than the previous one without requiring small IR_2 products. Still, operation of M_1 with large overdrives remains beneficial.

Mismatched mirrors are an additional error source in a great many analog circuits. Matching may be improved by using large devices biased at reasonably large overdrives. Systematic mismatch is reduced if the drain–source voltages are made as close to equal as possible. The circuit in Figure 10.12 uses another op-amp to force

operation of the PMOS mirror transistors with nominally equal drain-source voltages, thereby eliminating systematic errors due to channel-length modulation and DIBL.

With a constant-g_m circuit, one may reduce greatly the variation in g_m-dependent parameters – such as gain, input impedance, and noise figure of LNAs – as temperature, processing, and supply voltage vary.

10.7 SUMMARY

We have seen that the self-biased cell is quite versatile, permitting the generation of currents proportional to the ratio of a voltage in one branch to the resistance in the other. The voltage may be provided by a variety of elements, such as a forward-biased junction. Although a V_{BE} by itself has limited utility as a voltage reference because of its negative TC, its CTAT behavior is valuable in compensating the PTAT ΔV_{BE} in a bandgap reference circuit to yield an output roughly equal to V_{G0} with extremely small temperature variation. Even when parasitic bipolars are used in an otherwise CMOS circuit, the bandgap principle allows the synthesis of more accurate and stable voltages or currents than is possible with ordinary CMOS circuits.

Finally, a constant-g_m bias circuit was presented, allowing the stable biasing of such transconductance-sensitive circuits as filters and LNAs.

PROBLEM SET FOR VOLTAGE REFERENCES AND BIASING

PROBLEM 1   This problem explores in more detail the characteristics of the constant-g_m reference. Refer to Figure 10.13.

Rather than choosing $m = 4$, consider simply making m very large. In the limit, the transconductance approaches a value that is twice what is obtained when $m = 4$. Show this formally by deriving an expression for the transconductance of M_1 if the M_2 is S times as wide as M_1. To simplify the derivation, neglect body effect and assume that the PMOS mirror is an ideal 1 : 1 mirror.
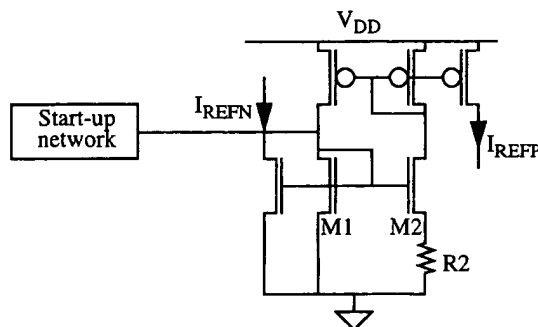

FIGURE 10.13. Basic constant-g_m reference.